Technical Recommendations

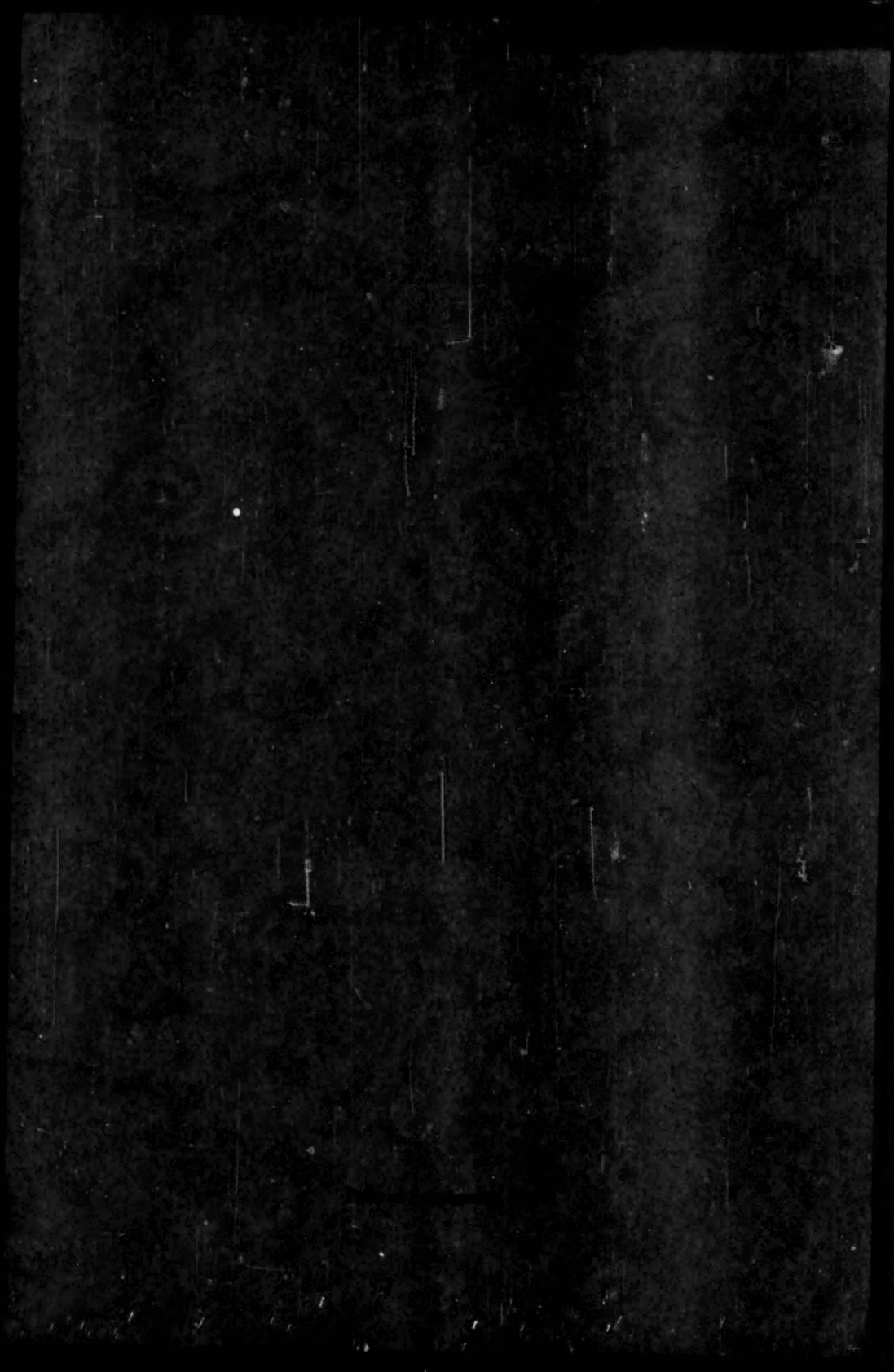
for Psychological Tests and Diagnostic Techniques

Vol. 51, No. 2, Part 2 March, 1954

Supplement to the

Psychological Bulletin

Published bimonthly by the
American Psychological Association



Technical Recommendations

for Psychological Tests and Diagnostic Techniques

Prepared by a joint committee of the American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education.

PUBLISHED BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1333 Sixteenth Street N.W., Washington 6, D. C.

Entered as second class mail matter at the post office at Washington, D.C., under the act of March 3, 1879. Additional entry at the post office at Menasha, Wisconsin. Acceptance for mailing at special rate of postage under the provisions of Sec. 34-40 Par. (D) provided for in Section 538, act of February 25, 1925, authorized August 6, 1947. Printed in U.S.A.

Foreword

This statement has been endorsed by the respective governing bodies of the American Psychological Association, the American Educational Research Association, and the National Council on Measurements Used in Education. The original drafts were developed in the APA Committee on Test Standards, whose members were Edward S. Bordin, R. C. Challman, H. S. Conrad, Lloyd G. Humphreys, Paul E. Meehl, Donald E. Super, and Lee J. Cronbach, chairman. The work was modified and extended in cooperation with the AERA committee (Jacob S. Orleans, chairman, Saul B. Sells, and J. R. Gerberich) together with three liaison members (Conrad, Cronbach, and Super) and the NCMUE committee, whose successive chairmen have been Gerberich, Henry Rinsland, and Robert L. Ebel. An extension of the recommendations to cover additional problems related to achievement tests is in preparation.

The statements presented here were submitted for criticism by specialists in test construction and use, including test publishers, and a preliminary version was published in the *American Psychologist* (*Amer. Psychologist*, 1952, **7**, 461–475) for wider examination. The present statement is the result of successive revisions.

Development and Scope of the Recommendations

Psychological and educational tests are used in arriving at decisions which may have great influence on the ultimate welfare of the persons tested, and of the community. Test users, therefore, wish to apply high standards of professional judgment in selecting and interpreting tests, and test producers wish to produce tests which can be of the greatest possible service. The test producer, in particular, has the task of providing sufficient information about each test so that users will know what reliance can safely be placed on it.

Professional workers agree that test manuals and associated aids to test usage should be made complete, comprehensible, and unambiguous, and for this reason there have always been informal "test standards." Publishers and authors of tests have adopted standards for themselves, and standards have been stated in textbooks and other publications. Through application of these standards, tests have attained a high degree of quality and usefulness.

Until this time, however, there has been no statement representing a consensus as to what information is most helpful to the test consumer. In the absence of such a guide, it is inevitable that some tests appear with less adequate supporting information than others of the same type, and that facts about a test which some users regard as indispensable have not been reported because they seemed relatively unimportant to the test producer. This report is the outcome of an attempt to survey the possible types of information that test producers might make available, to weigh the importance of these, and to make recommendations regarding test preparation and publication.

Improvement of testing has long been a concern of professional workers. In 1906, an APA committee, with Angell as chairman, was appointed to act as a general control committee on the subject of measurements. The purpose of their work was to standardize testing techniques, whereas the present effort is concerned with standards of reporting information about tests.

In a developing field, it is necessary to make sure that standardizing efforts do not stifle growth. The words of the earlier committee are appropriate today:

> The efforts of a standardizing committee are likely to be regarded with disfavor and apprehension in many quarters, on the ground that the time is not yet ripe for stereotyping either the test material or the procedure. It may be felt that what is called for, in the present immature condition of individual psychology, is rather the free invention and the appearance of as many variants as possible. Let very many tests be tried, each new investigator introducing his own modification; and then, the worthless will gradually be eliminated and the fittest will survive.

Issuing specifications for tests could indeed discourage the development of new types of tests. So many different sorts of tests are needed in present psychological practice that limiting the kind or the specifications would not be sound procedure. Appropriate standardization of tests and manuals, however, need not interfere with innovation. The recommendations presented here are intended to

assist test producers to bring out a wide variety of tests that will be suitable for all the different purposes for which tests should be used and to make those tests as valuable as possible.

Information Standards as a Guide to Producers and Users of Tests

The essential principle that sets the tone for this document is that a test manual should carry information sufficient to enable any qualified user to make sound judgments regarding the usefulness and interpretation of the test. This means that certain research is required prior to release of a test for general use by psychologists or school personnel. The results must be reported or summarized in the manual, and the manual must help the reader to interpret these results.

A manual is to be judged not merely by its literal truthfulness, but by the impression it leaves with the reader. If the typical professional user is likely to obtain an inaccurate impression of the test from the manual, the manual is poorly written. Ideally, manuals would be tested in the field by comparing the typical reader's conclusions with the judgment of experts regarding the test. In the absence of such trials, our recommendations are intended to apply to the spirit and tone of the manual as well as its literal statements.

A manual must often communicate information to many different groups. Many tests are used by classroom teachers or psychometrists with very limited training in testing. These users will not follow technical discussion or statistical information. At the other extreme of the group of readers, the available information about any test should be sufficiently complete for specialists in the area to judge the technical adequacy of the test. Sometimes the more technical information can be presented in a supplementary handbook, but it is most important that there be made available to the person concerned with the test a sound basis for whatever judgments his duties require.

The setting of numerical specifications has been avoided, even though it would have been tempting to say, for instance, that a validity coefficient ought to reach .50 before a test of Type A is ready for use or that a test of Type B should always have a reliability of .90 before it is used for the measurement of individual subjects. There are different problems in different situations, depending, for instance, on whether clinical analysis or personnel selection is involved, or whether preliminary or final decisions are being made. It is not appropriate to call for a particular level of validity and reliability, or to otherwise specify the nature of the test. It *is* appropriate to ask that the manual give the information necessary for the user to decide whether the accuracy, relevance, or standardization of the test makes it suitable for his purposes. These recommendations, then, suggest standards of *test description and reporting* without stating minimum statistical specifications.

The aim of the present standards is partly to make the requirements as to information accompanying published tests explicit and conveniently available. In arriving at those requirements, it has been necessary to judge what is presently the reasonable degree of compromise between pressures of cost and time, on the one hand, and

the ideal, on the other. The test producer ordinarily spends large sums of money in developing and standardizing a test. Insofar as these recommendations indicate the sort of information that would be most valuable to the people who use tests, test authors and publishers can then direct their funds to gathering and reporting those data. Validation on job criteria, for example, is essential before a vocational interest inventory can be used practically, but only a desirable addition for a values inventory, and irrelevant for an inventory designed to diagnose mental disorders. The recommendations therefore attempt to state what type of studies should be completed before a test is ready for release to the profession for operational use. The recommendations attempt to describe standards which are already reached by our better tests.

Tests to Which the Recommendations Apply

These recommendations cover not only tests as narrowly defined, but also most published devices for diagnosis and evaluation. The recommendations apply to interest inventories, personality inventories, projective instruments and related clinical techniques, tests of aptitude or ability, and achievement tests. The same general types of information are needed for all these varieties of tests. General recommendations have been prepared with all these techniques and instruments in mind. Since each type of test presents certain special requirements, additional comments have been made to indicate specific applications of the recommendations to particular techniques. Many principles of specific importance in measurement of achievement remain to be worked out in a subsequent statement.

Tests can be arranged according to degree of development. The highest degree of development is needed for tests distributed for use in practical situations where the user is unlikely to validate the tests for himself. Such a user must assume that the test does measure what it is presumed to measure on the basis of its title and manual. For instance, if a clerical aptitude measure is used in vocational guidance under the assumption that this will predict success in office jobs, there is very little possibility that the counselor could himself validate the test for the wide range of office jobs to which his clients might go.

At the other extreme of the continuum are tests in the very beginning stages of their development. At this point, perhaps the investigator is not sure whether his test is measuring any useful variable. Sometimes, because the theory for interpreting the test is undeveloped, the author restricts use of the test to situations where he himself knows the persons who will use the test, can personally caution them as to its limitations, and is using the research from these trials as a way of improving the test.

Between these tests which are so to speak embryonic, and the tests which are released for practical application without local validation, are tests released for somewhat restricted use. There are many tests which have been examined sufficiently to indicate that they will probably be useful tools for psychologists, but which are released with the expectation that the

user will conduct validation studies against performance criteria, or will verify suggested clinical interpretations by studying the subsequent behavior of persons in treatment. Examples are certain tests of spatial ability, and some inventories measuring such traits as introversion.

The present recommendations apply to devices which are distributed for use as a basis for practical judgments rather than solely for research. Most tests which are made available for use in schools, clinics, and industry are of this practical nature. Tests released for operational use should be prepared with the greatest care. They should be released to the general user only after their developer has gathered information which will permit the user to know for what use the test can be trusted. These statements regarding recommended information apply with especial force to tests distributed to users who have only that information about the test which is provided in the manual and other accessories. In the preparation of the recommendations, no attention was paid to tests which are privately distributed and circulated only to specially trained users. The recommendations also do not apply to tests presented in journal articles unless the article is intended to fulfill the functions of a manual.

A brief discussion of problems of projective techniques is needed here because of the opinion occasionally voiced that these devices are so unlike other testing procedures that they cannot be judged according to the same standards.

Many users of projective devices aim at idiographic analysis of an individual. Since this kind of analytical thinking places heavy reliance on the creative, artistic activity of the clinician, not all of this process can be covered in test standards. Thus, the recommendations herein presented are necessarily of a psychometric nature and should not be interpreted as meaning that projective techniques are intended primarily for such use. Nevertheless, proposals for arriving at such unique idiographic interpretations are almost always partially based upon some nomothetic premises, e.g., that a Rorschach determinant tends to correlate with a specified internal factor. There is no justification for failure to apply the usual standards in connection with these premises. Therefore, although these devices present unusual problems, the user of projective techniques requires much of the same information that is needed by users of other tests.

Even though the data from projective tests are more often qualitative than quantitative, these devices should be accompanied by appropriate evidence on validity, reliability, and so on. A projective test author need not identify his test's validity by correlating it with any simple criterion. But if he goes so far as to make any generalization about what "most people see" or what "schizophrenics rarely do," he is making an out-and-out statistical claim and should be held to the usual rules for backing it up. Obviously, when quantitative information is asked for in the recommendations, it is expected to apply where a quantitative kind of claim has been made. If a projective test makes no such claim, a recommendation would not be meaningful for it.

On the other hand, clinicians sometimes forget that the words "more," "usual," "typical," and the like are *quantity* words. Any textual discourse containing such words, or any verbal statement describing a correspondence between test performance and personality structure is making a quantitative claim. The only difference between such a verbal statement and a statistical table is the relative exactness of the latter. For this reason, many of the recommendations apply to aspects of projective instruments for which verbal rather than numerical interpretations are suggested.

The general topics to be covered in the recommendations are Dissemination of Information, Interpretation, Validity, Reliability, Administration, and Scales and Norms.

Many comments have been made to amplify and illustrate the recommendations. Tests mentioned in the comments have not been singled out as being particularly good or poor tests. The tests used for illustrative purposes were chosen because they are widely known, except where some less prominent test provides an unusually clear illustration of the point under discussion. These references to tests are not intended as critical evaluations of the test as a whole and should not be quoted or referred to in test advertising.

Three Levels of Recommendations

Manuals can never give all the information that might be desirable, because of economic limitations. At the same time, restricting this statement of recommendations to essential information might tend to discourage reporting of additional information. To avoid this, recommendations are grouped in three levels: ESSENTIAL, VERY DESIRABLE, and DESIRABLE. Each proposed requirement is judged in the light of its importance and the feasibility of attaining it.

The statements listed as ESSENTIAL are intended to be the consensus of present-day thinking as to what is normally required for operational use of a test. Any test presents some unique problems, and it is undesirable that standards should bind the producer of a novel test to an inappropriate procedure or form of reporting. The ESSENTIAL standards indicate what information will be genuinely needed for most tests in their usual applications. When a test producer fails to satisfy this need, he should do so only as a considered judgment. In any single test, there will be very few ESSENTIAL standards which do not apply.

If some type of ESSENTIAL information is not available on a given test, it is important to help the reader recognize that the research on the test is incomplete in this respect. A test manual can satisfy all the ESSENTIAL standards by clear statements of what research has and has not been done and by avoidance of misleading statements. It will not be necessary to perform much additional research to satisfy the standards, but only to discuss the test so that the reader fully understands what is known (and unknown) about it.

The category VERY DESIRABLE is used to draw attention to types of information which contribute greatly to the user's understanding of the test. They have not been listed as ESSENTIAL for various reasons. For

example, if it is very difficult to acquire information (e.g., long-term follow-up), it cannot always be expected to accompany the test. At times a closely reasoned minority opinion regards a type of information as unimportant. Such information is still very desirable, since many users wish it, but it is not classed as ESSENTIAL so long as its usefulness is debated.

The category DESIRABLE includes information which would be helpful, but less so than the ESSENTIAL and VERY DESIRABLE information. Test users welcome any information of this type the producer offers.

When a test is widely used, the producer has a greater responsibility for investigating it thoroughly and providing more extensive reports. The larger sale of such tests makes such research financially possible. Therefore the producer of a popular test can add more of the VERY DESIRABLE and DESIRABLE information in subsequent editions of the manual. For tests having limited sale, it is unreasonable to expect that as much of these two categories of information will be furnished. In making such facts available, the producer performs a service beyond the level that can reasonably be anticipated for most tests at this time.

The Audience for These Recommendations

These recommendations are intended to guide test development and reporting. A good deal of the information to be reported about tests is technical, and therefore the wording of the recommendations is of necessity technical. They should be meaningful to readers who have had a minimum of one substantial course in tests and measurements.

One audience for the recommendations is the authors and publishers who are responsible for test development. The recommendations should also aid the thinking of test users working either in psychology or education. It is not expected that the classroom teacher who has not had a course in tests and measurements will himself use this report. The report should, however, be helpful to directors of research, school psychologists, counselors, supervisors, and administrators who select tests to use for various school purposes.

As an aid to test development, the recommendations provide a kind of check list of factors to consider in designing standardization and validation studies. Test authors should refer to them in deciding what studies to perform on their tests and how to report them in their manuals. Test publishers will be able to use them in planning revision of their present tests. In considering proposed manuals, publishers can suggest to authors the types of information which need to be gathered in order to make the manual as serviceable as it should be. Because of the ease with which such claims could be misinterpreted, it would not be appropriate to state in a test manual that it "satisfies" or "follows" these Technical Recommendations. There would be no such objection to a statement that an author had "attempted to take into account or considered" these recommendations in preparing the manual.

Almost any test can be useful for some functions and in some situations. But even the best test can

have damaging consequences if used inappropriately. Therefore, ultimate responsibility for improvement of testing rests on the shoulders of test users. These recommendations should serve to extend the professional training of these users so that they will make better use of the information about tests and the tests themselves. The recommendations draw attention to recent developments in thinking about tests and test analysis. The report should serve as a reminder regarding features to be considered in choosing tests for a particular program.

Professional thinking about tests is much influenced by test reviews, textbooks on testing, and courses in measurement. These recommendations may be helpful in improving such aids, for instance, by suggesting features especially significant to examine in a test review. The recommendations can be a teaching aid in measurement courses. It is important to note that publication of superior information about tests by no means guarantees that tests will be used well. The continual improvement of courses which prepare test users and of leadership in all institutions using tests is a responsibility in which everyone must share.

Revision and Extension

For many reasons, it will be necessary to revise the recommendations periodically.[1] Despite the care with which the standards have been developed, experience will no doubt reveal that some of our judgments would benefit from further examination. New tests will present problems not considered in the present work. The improvement of statistical techniques and psychometric theory will yield better bases for test analysis. The efforts of test producers will lead to continued improvement in tests, and as this continues it will be possible to raise the standards so that the test user will have ever better information about his tools.

The recommendations here presented are intended to be used without reference to any enforcement machinery. The statement will be used by individual members of the professions to improve their own work.

[1]Continuing committees of the three associations are being established to receive comments on the recommendations and to plan appropriate revision. The 1953–1954 APA Committee on Test Standards consists of Edward S. Bordin, chairman, Paul E. Meehl, David Tiedeman, Jacob S. Orleans, *ex officio*, and R. L. Ebel, *ex officio*.

The Recommendations

A. Dissemination of Information

The test user needs information to help him select the test which is most adequate for a given purpose. He must rely in large part on the test producer for such data. The practices in furnishing the needed information have varied. In the case of some tests, the user has had access to virtually nothing beyond directions for administering and scoring the test, and norms of uncertain origin. On the other hand, other tests have manuals which furnish extensive data on the development of the test, its validity and reliability, the origin of the norms, the kinds of interpretations which are appropriate, and the uses for which it can be employed. The diversity of practice in making information about tests available suggests the need for standards for the dissemination of information.

A 1. When a test is published for operational use, it should be accompanied by a manual which takes cognizance of the detailed recommendations in this report. ESSENTIAL

[Comment: Sometimes information needed to support interpretations suggested in the manual cannot be presented at the time the manual is published. The manual satisfies the intent of recommendation A 1 if it points out the absence and importance of this information.

It should be recognized that a recommendation may not apply to a particular test. The manual writer "takes cognizance of" the recommendation if he examines it with care to make certain whether it has implications for his test. It is not proper to ignore a recommendation merely because the recommendation, while applicable to claims made for the test, is difficult to meet or has ordinarily not been met by similar tests.]

A 1.1 Some form of manual, presenting at least minimum information, should be given or sold to all purchasers of the test. ESSENTIAL

A 1.2 Where the information is too extensive to be fully reported in such a manual, the manual should summarize the ESSENTIAL information and indicate where further details may be found. ESSENTIAL

[Comment: The Differential Aptitude Tests provide an extensive manual, and also make further research data available through the American Documentation Institute. A great deal of the information about the Stanford-Binet is included in a book which all users must have. The Strong Vocational Interest Blank has been the subject of unusually thorough research which is reported in a technical book; a brief version of the ESSENTIAL information is given in a manual sold with the blanks.

For many projective techniques, such as the Rorschach and TAT, publications by persons other than the test author fulfill many functions of a manual. Insofar as a book about a technique fulfills the functions of a manual, the author has the same responsibility in preparing it as does the original author of the test.]

A 1.3 If information about the test is provided in a separate publication, any such publication should meet the same standards of accuracy as apply to the manual. ESSENTIAL

[Comment: A report in a professional journal, for instance, on the validity of an

instrument should meet the same standards of completeness and freedom from misleading impressions as a report in the manual. Recommendation A 1.3 applies also to advertising literature.]

A 2. The manual should be up-to-date. It should be revised at appropriate intervals. ESSENTIAL

[Comment: As criteria change, the predictive validity of a test may be altered. Also, the norms may require revision. Thus, a change in school objectives which places increased emphasis on problem solving in algebra, rather than on factoring and other mechanics, could appreciably affect the validity of an algebra aptitude test. It would also alter the norms for an algebra achievement test.]

A 2.1 When new information emerges, from investigations by the test author or others, which indicates that some facts and recommendations presented in the manual are substantially incorrect, a revised manual should be issued at the earliest feasible date. ESSENTIAL

[Comment: A revised manual for the Army Beta which arose out of World War I was issued in 1946. In contrast, although extensive published research points out the need for altering statements made in the manual of the Bernreuter Personality Inventory, no revised edition of that manual has been prepared. Likewise, the 1943 manual for the TAT has not been revised despite extensive development in the field since that date.]

A 2.2 When a test is revised or a new form is prepared, the manual should be thoroughly revised to take changes in the test into account. ESSENTIAL

[Comment: The Wechsler-Bellevue Scale was modified in several respects in the third edition of the manual. For example, the directions and scoring procedure were altered. The norms should have been reviewed or redetermined. Instead, the earlier tables for converting scores to IQ were carried over, without change, to the new edition.]

A 2.21 When a short form of a test is prepared by reducing the number of items or organizing a portion of the test into a separate form, new evidence should be obtained and reported for that new form of the test. VERY DESIRABLE

[Comment: This is especially important for inventories, where placing items in a new context might alter the person's responses. For example, the MMPI properly retains some items which were not scored in any key, because removing those items might alter the discriminating power of the items which were scored.]

A 2.22 When a short form is prepared from a test, the manual should present the correlation between the long and short forms, separately administered. DESIRABLE

A 2.3 The copyright date of the manual or the date of the latest revision should be clearly indicated. ESSENTIAL

B. Interpretation

In interpreting tests, the user always is responsible for making inferences as to the meaning and legitimate uses of test results. In making such judgments, he is dependent upon the available data about the test.

The degree to which a test manual can be expected to prepare the user for accurate interpretation and effective use of the test varies with the

type of test and the purpose for which it is used. For any test, it is sometimes necessary to make judgments which have not been substantiated by the published evidence. Thus the vocational counselor cannot expect to have regression equations available for predictions he must make from test scores, and the clinician must interpret a personality inventory on the basis of general data and theory because research on any one instrument is incomplete. The manual of a projective test cannot fully prepare the user for interpretation. Test users should be wary of interpreting projective test results without supervised training with that device and instruction in the clinical concepts and data which are part of its background.

This problem of accuracy is not the only consideration related to test interpretation. An equally important concern is the examinee's reactions to interpretations of his test scores, if the interpretation is made to him. Many educational and clinical uses of tests require reporting the interpretations to the person tested. The teacher who interprets the results of academic achievement tests affects the student's self concept and future learning. The clinician, in making interpretations which bear upon the client's areas of conflict, may unwittingly intensify those conflicts.

B 1. Insofar as possible, the test, the manual, record forms, and other accompanying material should assist users to make correct interpretations of the test results. ESSENTIAL

B 1.1 Names given to tests, and to scores within tests, should be chosen to minimize the risk of misinterpretation by test purchasers and subjects. ESSENTIAL

[Comment: The Army General Classification Test, the Blacky Test, and the Draw-A-Person Test are examples of names based on the content or process involved in the test which carry no unwarranted suggestions as to characteristics measured. Such names as "culture-free test," "primary abilities test," "measure of mental growth," and "temperament test" are likely to suggest interpretations going beyond the demonstrable meaning of test scores.

Names designed to disguise the purpose of a test from a subject may properly be used. In such a case, the manual should contain in an early and conspicuous place an explanation of the reason for choosing this name and a statement of what in fact the test is supposed to measure.

B 1.1 and subordinate recommendations can be followed in developing new tests, but it will rarely be feasible to rename established tests, even when this would be desirable.]

B 1.11 Interest and personality indices based on the self-report principle should be called "inventories," "questionnaires," or the like, rather than "tests." ESSENTIAL

B 1.2 The manual or other accompanying material should describe the process by which interpretations are to be derived from test scores. VERY DESIRABLE

[Comment: The manual need not include such information as all professionally qualified users may be expected to have. The original manual for the Differential Aptitude Tests presented a few profiles and gave an interpretation and a too brief case summary for each one. Later, more extensive case reports were reported in *Counseling from Profiles*, a supplementary booklet on the test, and the sketchy profiles were re-

moved from the manual. The case reports avoid oversimplification and emphasize the possible influence of non-test data on test interpretation.

The *Atlas for the MMPI* makes available for study examples of a variety of complex personality profiles. Few other personality inventories are supplemented by such materials as aids in their interpretation.]

B 1.21 The manual should draw the user's attention to data other than the test scores which need to be taken into account in interpreting the test. VERY DESIRABLE

[Comment: For example, Murray's TAT manual states that "the psychologist should know the following basic facts: the sex and age of the subject, whether his parents are dead or separated, the ages and sexes of his siblings, his vocational and his marital status."]

B 1.22 When case studies are used as illustrations for the interpretations of test scores, the examples presented should include some relatively complicated cases whose interpretation is not clear-cut. VERY DESIRABLE

B 1.23 Where a certain misinterpretation of a given test is known to be frequently made (or can reasonably be anticipated in the case of a new test), the manual should draw attention to this error and warn against it. ESSENTIAL

[Comment: Since the Terman-McNemar Test of Mental Ability reports scores in terms of a deviation IQ rather than a ratio IQ, it discusses at some length the fact that deviation IQ's do not have the same properties as ratio IQ's. Complete avoidance of the term IQ for deviation scores would be a more certain way to avoid confusion.

Another common misconception is that intelligence tests are measures of inherent native ability alone; it would be desirable for manuals of such tests to caution against this interpretation.

Manuals for interest measures should make clear, and urge counselors to stress to the client, the fact that interest does not imply ability and is only one factor to be considered in choosing among occupations. A desirable caution of this type is found in the Lee-Thorpe Occupational Interest Inventory.]

B 2. The test manual should state explicitly the purposes and applications for which the test is recommended. ESSENTIAL

B 2.1 If a test is intended for research use only, and is not distributed for operational use, that fact should be prominently stated in the accompanying materials. ESSENTIAL

[Comment: If, for example, an investigator plans to release tests developed by factor analysis for research use, it would be appropriate to print "distributed for research use only" on the test package or cover of the booklet of directions. This would serve to caution against premature use of the tests in guidance.]

B 3. The test manual should indicate the professional qualifications required to administer and interpret the test properly. ESSENTIAL

B 3.1 Where a test is recommended for a variety of purposes or types of inference, the manual should indicate the amount of training required for each use. ESSENTIAL

[Comment: One suggested categorization of tests approved by the APA is as follows:[2]

Level A. Tests or aids which can ade-

[2] APA Code of Standards for Test Distribution, *American Psychologist*, November, 1950. This statement also includes descriptions of general levels of training which correspond to the three levels of tests.

quately be administered, scored, and interpreted with the aid of the manual and a general orientation to the kind of organization in which one is working. (E.g., achievement or proficiency tests.)

Level B. Tests or aids which require some technical knowledge of test construction and use, and of supporting psychological and educational subjects such as statistics, individual differences, and psychology of adjustment, personnel psychology, and guidance. (E.g., aptitude tests, adjustment inventories with normal populations.)

Level C. Tests and aids which require substantial understanding of testing and supporting psychological subjects, together with supervised experience in the use of these devices. (E.g., projective tests, individual mental tests.)

The manual might identify a test according to one of the foregoing levels, or might employ some form of statement more suitable for that test. Regarding a particular industrial personnel test, the manual might say: "This test can be administered and scored by an intelligent clerical employee, but decisions regarding hiring and related interpretations should be made only by a psychologist or personnel manager who has studied fundamental statistics including correlation. Only a vocational counselor with specialized graduate training should use the test for vocational guidance."]

B 3.11 The manual should not imply that the test is "self-interpreting," or that it may be interpreted by a person lacking proper training. ESSENTIAL

B 3.12 The manual should point out the counseling responsibilities assumed when a tester communicates interpretations about ability or personality traits to the person tested. ESSENTIAL

[Comment: While examinees may properly score their own interest inventories and examine their own profiles, the Manual for the Kuder Preference Record properly recommends that they should make interpretations and future plans only with professional help in individual or group counseling situations.]

B 3.2 The manual should draw attention to references dealing with the test in question with which the user should become familiar before attempting to interpret the test. The statement should avoid the implication that this constitutes the only training needed, if other training is required. VERY DESIRABLE

B 4. When a test is issued in revised form, the nature and extent of any revision, and the comparability of data for the revised and the old test should be explicitly stated. ESSENTIAL

[Comment: An example of desirable practice is found in the manual for the revised edition of the Study of Values.]

B 5. Statements in the manual reporting relationships are by implication quantitative, and should be stated as precisely as the data permit. If data to support such a statement have not been collected, that fact should be made clear. ESSENTIAL

[Comment: Writers sometimes say, for example, "Spatial ability is required for architectural engineering" or, "Bizarre responses often indicate schizophrenic tendencies." Such statements need to be made more definite. In what proportion of cases giving bizarre responses does schizophrenia develop? How much does architectural success depend upon spatial ability? Numerical data would provide the needed answer.]

B 5.1 When the term "significant" is employed, the manual should make clear whether statistical or practical

significance is meant, and the practical significance of statistically reliable differences should be evaluated. ESSENTIAL

B 5.2 The manual should clearly differentiate between an interpretation justified regarding a group taken as a whole, and the application of such an interpretation to each individual within the group. ESSENTIAL

[Comment: For example, if the standard error of measurement is five points, this statement should not be presented so as to imply that the obtained score for any one individual is within five points of his true score. For a single pupil, the difference between the obtained and true score might be very much larger.]

C. Validity

Validity information indicates to the test user the degree to which the test is capable of achieving certain aims. Tests are used for several types of judgment, and for each type of judgment, a somewhat different type of validation is involved. We may distinguish four aims of testing:

1. The test user wishes to determine how an individual would perform at present in a given universe of situations of which the test situation constitutes a sample.

2. The test user wishes to predict an individual's future performance (on the test or on some external variable).

3. The test user wishes to estimate an individual's present status on some variable external to the test.

4. The test user wishes to infer the degree to which the individual possesses some trait or quality (construct) presumed to be reflected in the test performance.

Thus, a vocabulary test might be used simply as a measure of present vocabulary, as a predictor of college success, as a means of discriminating schizophrenics from organics, or as a means of making inferences about "intellectual capacity."

Four Types of Validity

To determine how suitable a test is for each of these uses, it is necessary to gather the appropriate sort of validity information. These four aspects of validity may be named content validity, predictive validity, concurrent validity, and construct validity.

a. Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn. Content validity is especially important in the case of achievement and proficiency measures.

In most classes of situations measured by tests, quantitative evidence of content validity is not feasible. However, the test producer should indicate the basis for claiming adequacy of sampling or representativeness of the test content in relation to the universe of items adopted for reference.

b. Predictive validity is evaluated by showing how well predictions made from the test are confirmed by evidence gathered at some subsequent time. The most common means of checking predictive validity is correlating test scores with a subsequent criterion measure. Predictive uses of tests include long-range prediction of intelligence measures, prediction of vocational success, and prediction of reaction to therapy.

c. Concurrent validity is evaluated by showing how well test scores correspond to measures of concurrent criterion performance or status. Studies which determine whether a test discriminates between presently identifiable groups are concerned with concurrent validity. Concurrent validity and predictive validity are quite similar save for the time at which the criterion is obtained. Among the problems for which concurrent validation is used are the validation of psychiatric screening instruments against estimates of adjustment made in a psychiatric interview, differentiation of vocational groups, and classification of patients. It should be noted that a test having concurrent validity may not have predictive validity.

d. Construct validity is evaluated by investigating what psychological qualities a test measures, i.e., by demonstrating that certain explanatory constructs account to some degree for performance on the test. To examine construct validity requires both logical and empirical attack. Essentially, in studies of construct validity we are validating the theory underlying the test. The validation procedure involves two steps. First, the investigator inquires: From this theory, what predictions would we make regarding the variation of scores from person to person or occasion to occasion? Second, he gathers data to confirm these predictions.

There are various specific procedures for gathering data on construct validity. If it is supposed that form perception on the Rorschach test indicates probable ability to resist stress, this supposition may be validated by placing individuals in an experimental stress situation and observing whether behavior corresponds to prediction. Another much simpler procedure for investigating what a test measures is to correlate it with other measures; we would expect a valid test of numerical reasoning, for example, to be substantially correlated with other numerical tests, but not to be correlated with a clerical perception test. Factor analysis is another way of organizing data about construct validity.

We can distinguish among the four types of validity by noting that each involves a different emphasis on the criterion. In predictive or concurrent validity, the criterion behavior is of concern to the tester, and he may have no concern whatsoever with the type of behavior exhibited in the test. (An employer does not care if a worker can manipulate blocks, but the score on the block test may predict something he cares about.) Content validity is studied when the tester *is* concerned with the type of behavior involved in the test performance. Indeed, if the test is a work sample, the behavior represented in the test may be an end in itself. Construct validity is ordinarily studied when the tester has no definitive criterion measure of the quality with which he is concerned, and must use indirect measures to validate the theory. Here the trait or quality underlying the test is of central importance, rather than either the test behavior or the scores on the criteria.

It is ordinarily necessary to evaluate construct validity by integrating evidence from many different sources. The problem of construct validation becomes especially acute in the clini-

cal field since for many of the constructs dealt with it is not a question of finding an imperfect criterion but of finding any criterion at all. The psychologist interested in construct validity for clinical devices is concerned with making an estimate of a hypothetical internal process, factor, system, structure, or state and cannot expect to find a clear unitary behavioral criterion. Concern for validity is in no way a challenge to the dictum that prediction of behavior is the final test of any theoretical construction. But it is necessary to understand that *behavior-relevance* in a construct is not logically the same as *behavior-equivalence.* It is one thing to insist that in order to be admissible, a complex psychological construct must have some relevance to behavioral indicators; it is quite another thing to require that any admissible psychological construct must be *equivalent to* any direct operational behavior measure. Any position that cuts the test inference off from all possible nontest sources of confirmation appears to be an unreasonable one. If the test is to be interpreted in terms of internal constructs, there must be some facts, quantitative or not, that would argue for the existence of the particular internal system postulated. An attempt to identify any one criterion measure or any composite as *the* criterion aimed at is, however, usually unwarranted.

This viewpoint, while fraught with grave dangers and sometimes misused, is nevertheless methodologically sound. The clinician interested in construct validity has in mind an admittedly incomplete construct, the evidence for which is to be found roughly in such-and-such behavioral domains. The vagueness of the construct is an inevitable consequence of the incompleteness of current psychological theory, and cannot be rectified faster than theory grows and is confirmed. At a given stage of theoretical development, the only kind of prediction that can be made may be that certain correlations should be positive, or that patients who fail to conform to a group trend should be expected with considerable frequency to exhibit such-and-such an additional feature, or the like. It is clear that these deductions do involve behavioral prediction. They require the test-constructs to be behaviorally relevant. But they still do not necessarily *identify* any of the test-inferred constructs or variables with any criterion measure. A clinician may say, "I expect to find cases of psychosomatic ulcer showing large discrepancies between latent n Succorance as inferred from TAT stories and manifest n Succorance as revealed by the score on a questionnaire." Such a declaration leads to an empirical test.

The correlation or measure of discrimination obtained in studying construct validity is not to be taken as the "validity coefficient," in the same sense that prediction of washouts during flight training is *the* validity coefficient for the battery employed. Studies of many such predictions, possibly involving quite independent components of theory, will in the mass confirm or disconfirm the claims made.

One tends to ask regarding construct validity just what *is* being validated—the test or the underlying hypothesis? The answer is, *both*, simultaneously. If one predicts an empirical relation by supposing a cer-

tain personality organization, the verification of this prediction tends to confirm both the component suppositions that gave rise to it. True, there might be plausible alternative hypotheses, but this is always the case in science. The more alternatives there are, the more cumulated evidence is needed to justify confidence in the particular test-hypothesis pair. A further characteristic of this type of validity inference is that the construct itself undergoes modification as evidence accumulates. We do not merely alter our confidence in the correctness of the construct, or in the estimates of its magnitudes, but we actually reformulate or clarify our characterization of its nature on the basis of new data.

It must be kept in mind that these four aspects of validity are not all discrete and that a complete presentation about a test may involve information about all types of validity. A first step in the preparation of a predictive instrument may be to consider what constructs or predictive dimensions are likely to give the best prediction. Examining content validity may also be an early step in producing a test whose predictive validity is ultimately of major concern. Even after satisfactory predictive validity has been established, information relative to construct validity may make the test more useful. To analyze construct validity, our total background of knowledge regarding validity would be brought to bear.

Application of the Concepts to Ability Tests

Several examples of the application of these principles to intelligence tests should clarify the concepts involved. Correlations between an intelligence test used to select university students and later academic success are predictive validities. Such correlations will typically vary in size from those with criteria of proficiency in art or music at the lower end to those with grades in science at the upper end. If the test is used to predict an art criterion, then the correlation obtained, even though low, is the predictive validity of the test. Even if validities of intelligence tests are corrected for attenuation, a value substantially less than unity is the usual result. This is not interpreted, when predictive validity is at issue, that the criterion is an imperfect index of intelligence. Rather the test is regarded as an imperfect index of the criterion.

Relationships of subscores on an intelligence test to membership in various clinical groups are an example of evidence concerning concurrent validity. Again, low or imperfect validities are interpreted as due to inadequacies in the test as a discriminating device. A test is likely to be developed for making discriminations if it is difficult to measure status on a criterion directly. If the direct measurement of the criterion is expensive, dangerous, or highly unreliable, tests having concurrent validity are needed to assess status on the criterion indirectly.

Content validity is indicated by a description of the universe of items from which selection was made, including a description of the selection process. The universe of items in intelligence test construction is usually defined by the types of items used originally by Binet. Judges' ratings of appropriateness of items

are frequently involved. Content validity is ordinarily of little direct interest to the user of intelligence tests. The distinction between verbal and nonverbal tests of intelligence is, however, based on content analysis.

Construct validity may be judged from all of the information ordinarily subsumed under the preceding categories. Certain types of information, however, are employed here alone. Examples are as follows: correlations with other tests of intelligence, correlations with ratings of intelligence, factor analyses, nature-nurture studies, and studies of the effects of practice upon test scores. All relate to the problem of the meaning of the concept of intelligence. From this point of view a low correlation of the test with athletic ability may be just as important and encouraging as a high correlation with reading comprehension. This reverses the earlier emphasis from the viewpoint of concurrent or predictive validity, where a low correlation indicated weakness in the test.

Information concerning construct validity is of help to the theorist in formulating hypotheses concerning individual differences and to the test constructor in improving intelligence tests. For the practical test user this information is most frequently used to generalize beyond established predictive and concurrent validities. The careful verification of theory should serve to reduce the errors of extrapolation, but does not reduce the necessity of objective check upon extrapolations whenever possible.

Application of the Concepts to Personality Inventories

Evidence of the predictive validity of personality questionnaires provides the basis for their use for screening. One of the screening uses is to identify persons who will become maladjusted (as in the armed services). If personality instruments are used as a basis for predicting vocational or educational achievement, this inference also rests upon *predictive* validity.

Evidence of *concurrent* validity supports the use of personality questionnaires for screening and diagnostic purposes. An example is the use of check lists to determine which students are presently most in need of counseling.

Interpretation of responses as self-description (e.g., by judging conservatism from responses to a group of statements) represents one kind of assumption of *content* validity in the context of personality inventories.

Construct validity is involved when the personality inventory is used to ascertain the personality traits or structure of the individual.

Predictive or concurrent validation of personality questionnaires can depend upon fairly clear-cut operational criteria, e.g., reporting for sick call, membership in one occupational group as compared to another, psychiatric classifications. On the other hand, in validation of conceptual inferences, problems arise because of the lack of a simple relationship between personality traits and overt behavior. The "retiring" person may not actually behave in an unsociable manner, but the social activities in which he engages may be less satisfying to him, and participating in them may result in emotional stress which manifests itself in tics or other psychosomatic symptoms. This type of validity will not be judged by the size of any given relationship be-

tween a score and one criterion, but by the pattern of relationships which have been demonstrated to hold between a score and a number of different kinds of behavior criteria.

Application of the Concepts to Interest Inventories

Most interest inventories are used for predictive purposes. In counseling, scores are discussed in the context of a consideration of educational or vocational plans of the client. Even if the counselor makes very restricted interpretations (e.g., "the number of preferences for mechanical activities you reported is exceeded by only 5 per cent of high school seniors"), the context in which this discussion occurs implies that this information has some direct bearing on future performance. The test is really interpreted as indicating something about the client's probable success, satisfaction, or continuity in some activity.

Description of the individual is a second use of interest inventories. Interests are described in terms of categories or traits. In some devices, the description of interests involves such broad categories as to be essentially a description of generalized personality traits. This involves content validity, and often construct validity.

Different inferences must be supported by different types of evidence. In general, since counseling involves consideration of a very large number of vocations, it is not expected that every judgment for which an interest inventory is used will be validated by direct empirical evidence. Clients wish to consider very many occupations and activities. It is not possible to perform empirical studies with respect to all these, and reasonable tentative inferences may often be made in the absence of evidence from empirical studies. Knowledge from internal analysis of the inventory, job descriptions, and other sources may permit interpretations that will assist the client. Such extrapolations should be made tentatively, however. Extrapolation is found in the use of the Strong Blank to describe such traits as "interest in social uplift occupations," and in the use of Kuder scores to describe interest in vocations for which validity has not been tested.

Application of the Concepts to Projective Techniques and Related Clinical Methods

Predictive, concurrent, and construct validity all have pertinence to projective techniques although construct validity greatly overshadows the other two kinds. The prediction of a specific act of behavior is rarely made on the basis of projective instruments. Even the prediction of less specific behavior, such as "ability to profit from psychotherapy" is seldom made on the basis of projective techniques alone; in fact, there are a number of workers in this field who take the position that such an attempt should never be made.

Concurrent validity may be desired in projective techniques and clinical use of ability tests since they are used in making diagnostic classifications.

C 1. When validity is reported, the manual should indicate clearly what type of validity is referred to. The unqualified term "validity" should be

avoided unless its meaning is clear from the context. ESSENTIAL

[Comment: The manual should make clear what type of inference the validation study reports. No manual should report that "this test is valid." In the past, evidence that is not appropriately termed evidence of validity has been presented in the manual under that heading. For example, the "validity" report of the Thurstone Interest Schedule deals solely with item-test correlations. The discussion of item-test correlations in the manual of the Heston Personal Adjustment Inventory illustrates how such data may be used in reporting test validity without risk of misleading readers.

It is not desirable for the manual to state that any one type of evidence is the only possible sort of validity evidence. The following statement made regarding the Ohio Penal Classification Test is misleading: "The only criterion which establishes an intelligence test as valid is that labelled 'expert judgment' or 'expert agreement.' "]

C 2. The manual should report the validity of each type of inference for which a test is recommended. If validity of some recommended interpretation has not been tested, that fact should be made clear. ESSENTIAL

[Comment: In a test used for guidance it is obviously impossible to present predictive validities for all possible criteria in which a counselor might be interested. The manual should make clear to the test user the nature and extent of the extrapolations suggested by the author of the test, or forced upon him by the problem confronting him. Enough information is available concerning intelligence tests, for example, that the limits of generalization can be fairly accurately gauged. Less is known about tests of spatial ability and they cannot be readily applied as predictors for criteria for which validity studies have not been made. Hazardous extrapolation is likewise involved when tests are suggested as predictors of jobs solely on the basis of job analysis information.]

C 2.1 The manual should indicate which, if any, of the interpretations usually attempted for tests such as the one under discussion have not been substantiated or are based merely on clinical impressions. ESSENTIAL

[Comment: An example of a highly desirable practice is the warning to readers in the manual of the Purdue Pegboard: "Generalizations concerning the validity of any test should be made with great caution, and this is particularly true of dexterity tests. As Seashore has reported, motor skills are quite specific and ordinarily not highly correlated with each other. This situation perhaps accounts for the fact that a given dexterity test may have a rather satisfactory validity for certain manipulative jobs and yet be unsuitable for other manipulative jobs which might seem to be very similar. It is therefore highly desirable to conduct a study of the validity of the several Pegboard tests among employees on specific jobs for which the use of the test is contemplated, rather than attempt to generalize from available validity studies."]

C 2.11 If the manual for an inventory suggests that the user consult specific items as a basis for personality assessment, it should either present validation data for this use or call attention to their absence. The manual should also warn of the wide margins of error inherent in such interpretative procedures. ESSENTIAL

C 2.12 Validity of self-report as a description of the person's behavior can be demonstrated only by comparing responses on single items to

observed behavior. In the absence of such evidence, the manual should warn the reader that such references are subject to extreme error and should be used only to direct further inquiry, as in a counseling interview. ESSENTIAL

[Comment: If two investigators using similar criteria obtain very different predictive validities for a test, a presentation of both sets of facts in the test manual is in order. If a test of mechanical comprehension is validated against a clerical criterion, on the other hand, there is probably no value in reviewing these data in the manual. Badly controlled or badly analyzed studies need not be reported in the manual.

Validation samples are frequently small, with large standard errors of resulting coefficients. The only way in which large samples can be built is to pool results from several comparable studies. The cumulation of validating studies serves to set the limits on generalization, by demonstrating whether a test applies equally well in a variety of situations. Desirable practice is illustrated by the summary of validation studies provided in the 1946 manual for the Minnesota Clerical Test.]

Content Validity

C 3. Findings based on logical analysis should be carefully distinguished from conclusions established by correlation of test behavior with criterion behavior. ESSENTIAL

[Comment: Content validity may be established by demonstrating that a test samples a particular area. The user cannot judge, from this alone, how well the test permits drawing conclusions about any form of behavior other than the test behavior. For instance, it is reported that an occupational interest inventory inquires about a sample of items, chosen to represent vocational areas according to their frequency of occurrence. This is important information about the content validity of the interest scores, but it does not alone establish whether the student's scores predict how well he will be satisfied in a given type of job.]

C 4. If a test performance is to be interpreted as a sample of performance in some universe of situations, the manual should indicate clearly what universe is represented and how adequate the sampling is. ESSENTIAL

C 4.1 The universe of content should be defined in terms of the sources from which items were drawn, or the content criteria used to include and exclude items. ESSENTIAL

[Comment: For example, the manual for the Lee-Thorpe Occupational Interest Inventory describes the method used in devising items from the definitions in the Dictionary of Occupational Titles.]

C 4.2 The method of sampling items within the universe should be described. ESSENTIAL

[Comment: R. H. Seashore prepared a vocabulary test, defining his universe as all words in a certain unabridged dictionary, and sampled according to a definite plan.]

C 4.3 If items are regarded as a sample from a universe, a coefficient of internal consistency should be reported for each descriptive score, to demonstrate the extent to which the score is saturated with common factors. ESSENTIAL

[Comment: The present Lee-Thorpe manual does not report the internal consistency of its scales. See additional recommendations D 5–D 6 regarding internal consistency studies.]

C 4.4 If test performance is to be interpreted as a sample of performance in some universe of situations,

and if the test is administered with a time limit, evidence should be presented concerning the effect of speed on test scores. ESSENTIAL

[Comment: The most satisfactory evidence would be the correlation of one form, given with the usual time limit, against another form given with unlimited time. This could be compared to the form-form coefficient with time limits on both forms. Other simpler information about degree of speeding should be given when this correlational study is impractical.]

C 4.5 The date at which any study of the adequacy of sampling was made should be reported, and also the date of any sources of items. ESSENTIAL

[Comment: In achievement testing, it is frequently the practice to select items by a careful sampling from textbooks to identify significant topics. Textbooks and courses of study change, however, and the test which was once an excellent sample becomes obsolete. Therefore the manual should report some such statement as the median copyright date of the textbooks studied, or the date at which the experts agreed that the items were representative. In another field, the Mooney Problem Checklist lists problems which are common to students, on which each individual is to check those which concern him. The Mooney manual properly reports the date when the list was collected. After this list has been used for many years, it will be valuable to conduct a further study to determine whether student problems have changed significantly, and, if so, to change the test and manual accordingly.]

Predictive Validity

C 5. When predictive validity is determined by statistical analysis, the analysis should be reported in a form from which the reader can determine confidence limits of estimates regarding individuals, or the probability of misclassification of the individual on the criterion. ESSENTIAL

C 5.1 Statistical procedures which are well known and readily interpreted should be used in reporting validity whenever they are appropriate to the data under examination. Any uncommon statistical techniques should be explained. ESSENTIAL

C 5.11 Reports of statistical validation studies should ordinarily be expressed by: (*a*) correlation coefficients of familiar types; (*b*) description of the efficiency with which the test separates groups, indicating amount of misclassification or overlapping; or (*c*) expectancy tables. ESSENTIAL

[Comment: Reports of differences between means of groups, or critical ratios, are by themselves inadequate information regarding predictive validity. If a sample is large, high critical ratios may be found even when classification is very inaccurate.

In general, since manuals are directed to readers who have limited statistical knowledge, every effort should be made to communicate validity information clearly. An example of unwise use of a novel statistical method is found in the manual for the Ohio Penal Classification Test. Ten cases were chosen, separated at five-point intervals along the OPCT IQ scale. The IQ's were then correlated with Wechsler IQ's, yielding a rank correlation of .93. This correlation is greater than would be obtained for any sample not artificially spread along the scale. While unusual statistical procedures should be used for special problems, they should not be used where standard methods are equally or more efficient for evaluating the data. They certainly should be presented so that they will not mislead the typical user of the manual.

When a test is recommended for the purpose of dividing patients among discrete categories, correlational measures of association should be supplemented by percentage figures on misclassification, i.e., "false positives" and "false negatives." When validation involves comparison of men in an occupation with men-in-general, the comparison should be presented in such a way as to make clear the degree to which the occupational group overlaps the general group.]

C 5.2 An over-all validity coefficient should be supplemented with evidence as to the validity of the test at different points along the range, unless the author reports that the validity is essentially constant throughout. VERY DESIRABLE

[Comment: This might be reported by giving the standard error of estimate at various test score levels, or by indicating the proportion of hits, misses, and false positives at various cutting scores. The Metropolitan Reading Readiness Test reports the number of failures in primary reading expected at each level of test score.]

C 5.3 Test manuals should not report coefficients corrected for unreliability of the test as estimates of predictive validity. ESSENTIAL

[Comment: Corrections for attenuation are very much open to misinterpretation, and if misinterpreted give an unjustifiably favorable picture of the validity of the test. The hazard is illustrated in the manual for the Heston Personal Adjustment Inventory. Heston reports correlations between inventory scores and criterion ratings, and also reports the correlations augmented to correct for attenuation. He then applies significance tests to the augmented correlations rather than to the raw correlations only. Further, he comments that the augmented correlations "are as high as those often secured between college aptitude tests and college grades." This comparison is improper, since Heston is comparing his augmented coefficients with uncorrected coefficients for ability tests.]

C 5.31 If such coefficients are reported for the special purpose of studying construct validity, the uncorrected coefficients must be reported also and the proper interpretation of the corrected coefficients must be discussed. ESSENTIAL

C 6. All measures of criteria should be described accurately and in detail. The manual should evaluate the adequacy of the criterion. It should draw attention to significant aspects of performance which the criterion measure does not reflect and to the irrelevant factors which it may reflect. ESSENTIAL

[Comment: Desirable practices are illustrated in the manual of the General Clerical Test, where validity is reported in three specific studies. The nature of the criterion, and the nature of the work done by the employees tested is described. Limitations on the data are mentioned, and stress is placed on the necessity of making comparable studies with local criteria in any new situation where the test is to be applied.

For specific types of criteria, particular cautions in description are needed to avoid misconceptions or ambiguities. Some of these are listed in the recommendations which follow.]

C 6.1 When validity of a test is measured by agreement with psychiatric diagnoses, the diagnostic terms should be specific and the categories clearly described. VERY DESIRABLE

[Comment: "Paranoid schizophrenia, chronic" is preferable as a category to "schizophrenia." Since the types of patients included in specific diagnostic

classifications vary to some extent depending on the point of view of the psychiatrists, a description of each diagnostic category used in the validity study should be presented. An example of good practice is found in Rapaport's *Diagnostic Psychological Testing* where each diagnostic group is summarily described in terms of characteristics judged by the psychiatrists to be basic.]

C 6.11 If the individual usage given to a vague or variable clinical term by the validating psychiatrist is not known, this fact should be clearly stated and the reader warned that other raters or measuring devices might not agree with the criterion. VERY DESIRABLE

C 6.12 When validity of a clinical test is indicated by agreement with psychiatric judgment, the training, experience, and professional status (e.g., diplomate) of the psychiatrist should be stated. VERY DESIRABLE

C 6.13 When validity of a clinical test is indicated by agreement with psychiatric judgment, the amount and character of the patient contacts upon which the judgment is based should be stated. ESSENTIAL

C 6.2 When validity of an aptitude test is determined for predicting performance in an occupation, the occupation should be accurately defined. The test user should be given a clear understanding as to what duties are performed by workers in that occupation. ESSENTIAL

C 6.21 Where a wide range of duties is subsumed under a given occupational label, the test user should be warned against assuming that only one pattern of interests or abilities can be satisfied in the occupation. VERY DESIRABLE

C 6.3 When validity of an aptitude or interest test for predicting performance in a course or curriculum is reported, the character of the course or curriculum should be clearly defined. The test user should be given a clear understanding as to what types of performance are required in the course. ESSENTIAL

C 6.4 When predictive validity of an interest test is reported, the manual should state whether the criterion indicates satisfaction, success, or merely continuance in the activity under examination. ESSENTIAL

[Comment: When validation data compare men in an occupation to men-in-general, the manual should point out the limitations of presence in an occupation as a sign of success.]

C 6.5 The time elapsing between the test and determination of the criterion should be reported. ESSENTIAL

C 6.51 If a test is recommended for long-term predictions, but data from longitudinal studies are not presented, the manual should emphasize that predictions of this sort have uncertain validity. ESSENTIAL

C 7. The reliability of the criterion should be reported if it can be determined. If such evidence is not available, the author should discuss the probable reliability as judged from indirect evidence. VERY DESIRABLE

[Comment: When validity is measured by agreement of the test with psychiatric judgment, for example, statistical evaluation of the agreement among judges should be reported.]

C 7.1 If validity coefficients are corrected for unreliability *of the criterion*, both corrected and uncorrected coefficients should be reported and properly interpreted. ESSENTIAL

C 8. The date when validation data were gathered should be reported. ESSENTIAL

C 8.1 If the criterion, the conditions of work, the type of person likely to be tested, or the meaning of the test items is suspected of changing materially with the passage of time, the validity of the test should be rechecked periodically and the results reported in subsequent editions of the manual. VERY DESIRABLE

[Comment: Criterion data for the Psychologist scale of the Strong Vocational Interest Blank were gathered in 1927. Subsequent research showed that these psychologists were no longer representative of the field. The current manual reports the date (1948) of the validating studies for the revised key.]

C 9. The criterion score of a person should be determined independently of his test score. The manual should describe precautions taken to avoid contamination of the criterion or should warn the reader of any possible contamination. ESSENTIAL

C 9.1 When the criterion consists of a rating, grade, or classification assigned by an employer, teacher, psychiatrist, etc., the manual must state whether the test data were available to the rater or were capable of influencing his judgment in any way, e.g., indirectly through other reports of the psychologist. ESSENTIAL

C 9.11 If the test data could have influenced the criterion rating, this fact should be emphasized and the user warned that the reported validities are thus contaminated and are likely to be spuriously raised. ESSENTIAL

C 10. Test scores to be used in validation should be determined independently of criterion scores. ESSENTIAL

[Comment: In any test where knowledge about the subject may influence test administration or scoring, for instance in individual intelligence tests or projective techniques, the test administrator should possess no knowledge of the behavior of the subject outside the test situation. The manual should discuss the extent to which contamination of this type is possible unless it is obvious from the character of the test that no such contamination could occur. Recommendation C 11 below refers to a special kind of contamination frequently found in studies of objective tests.]

C 11. When items are selected or a scoring key is established empirically on the basis of evidence gathered on a particular sample, the manual should not report validity coefficients computed on this sample, or on a group which includes any of this sample. The reported validity coefficients should be based on a cross-validation sample. ESSENTIAL

C 11.1 If the manual recommends certain regression weights, any validity reported for the composite should be based on a cross-validation sample. VERY DESIRABLE

[Comment: A possible exception to recommendation C 11.1 is that a cross-validation sample would not be required if an appropriate correction for shrinkage could be applied to data from the original sample. Corrections available at present are not adequate for this purpose.]

C 12. If the manual recommends that interpretation be based on the test profile, evidence should be provided that the shape of the profile is a valid predictor. VERY DESIRABLE

[Comment: One suitable method, for example, is to tabulate test profiles hav-

ing the same two highest scores, to show what proportion of these persons are successful or unsuccessful, and to compare the discriminating ability of these combined scores with that of a single score.]

C 12.1 If the interpretation emphasizes complex nuances of the profile pattern which cannot be fully specified and depend upon the clinical experiences of the user, evidence, specifying the training and experience of the clinicians, should be presented to show how much increase in accuracy over more simplified interpretations is gained. ESSENTIAL

C 12.2 If the matching method is used to establish validity for the test report as a whole, the manual should point out that this analysis does not establish the validity of the component variables. ESSENTIAL

C 13. The validation sample should be described sufficiently for the user to know whether the persons he tests may properly be regarded as represented by the sample on which validation was based. ESSENTIAL

C 13.1 The user should be warned against assuming validity when the test is applied to persons unlike those in the validating sample. ESSENTIAL

C 13.2 Appropriate measures of central tendency and variability of test scores for the validation sample should be reported. ESSENTIAL

C 13.3 The number of cases in the validation sample should be reported. The group should be described in terms of those variables known to be related to the quality tested: these will normally include age, sex, socioeconomic status, and level of education. Any selective factor which restricts or enlarges the variability of the sample should be indicated. ESSENTIAL

[Comment: In tests validated on patients, the diagnoses of the patients would usually be important to report. The severity or obviousness of the diagnosed condition should be stated when feasible. In tests for industrial use or vocational guidance, occupation and experience of the validation sample should be described.]

C 13.4 If the validation sample is made up merely of "available records," this fact should be stated. The test user should be warned that the group is not a systematic sample of any specifiable population. ESSENTIAL

C 13.5 A sample made up of "available records" should be discussed in some detail as to probable selective factors and their presumed influence on test variables. VERY DESIRABLE

C 13.6 If validation is demonstrated by comparing groups which differ on the criterion, the manual should report whether and how much the groups differ on other relevant variables. ESSENTIAL

[Comment: Groups which differ on a criterion may also differ in other respects, so that the test may be discriminating on a quality other than that intended. Score differences between types of patients, for instance, may reflect differences in age, education, or length of time in hospital, unless these factors are controlled.]

C 14. The author should base validation studies on samples comparable, in terms of selection of cases and conditions of testing, to the groups to whom the manual recommends that the test be applied. VERY DESIRABLE

C 14.1 If the test score distribu-

tion of the validation sample is markedly different from the distribution of the group with whom the test is ordinarily to be used, coefficients or other measures of discrimination should be corrected to the value estimated for the group to whom the test is to be given. ESSENTIAL

[Comment: A biserial correlation between a scholastic aptitude test and college success, where the persons distinguished are dropouts and honor students, will be much higher than a coefficient based on all entering students. The test will normally be applied to the latter group, and the validity coefficient should emphasize the power of the test in that group. A correction to raise the validity coefficient may likewise be needed when a test is validated on a group of selected employees. It is always preferable, however, to gather criterion data for an unselected group.]

C 14.2 In reporting coefficients corrected for range, the manual should report the original coefficient, and the distribution characteristics used in making the correction and the formula employed in making the correction. ESSENTIAL

C 14.3 Validation of tests intended for use in guidance should generally be based upon subjects tested at the time when they are making educational or vocational choices. VERY DESIRABLE

[Comment: Strong standardized his Vocational Interest Blank on men who were currently employed in the occupation in question. The ability of these scales to differentiate between occupational groups did not, in and of itself, warrant using the inventory in the counseling of high school or college students. Strong obtained better evidence by administering the inventory to students and ascertaining the nature of their later employment, thus establishing the relationship between preoccupational score and later occupation.]

C 14.4 If a test is presented as being useful in the differential diagnosis of patients, it should include evidence of the test's ability to separate diagnostic groups from one another. Emphasis should be placed on this rather than on the differentiation of diagnosed abnormal cases from the normal population. ESSENTIAL

C 15. If the validity of the test can reasonably be expected to be different in subgroups which can be identified when the test is given, the manual should report the validity for each group separately or should report that no difference was found. VERY DESIRABLE

C 15.1 Occupational predictions by means of interest tests should be validated within a group all of whom have the same stated vocational aim. DESIRABLE

[Comment: An interest inventory is an attempt to obtain more accurate and complete information than would be obtained by a simple question such as "List your preferred occupation." Whether the inventory yields useful information can be demonstrated only by showing that, among persons who give the same answer to this simple question, the test makes valid discriminations. It is important to move in the direction of reporting whether *among students stating a preference for engineering* (for example), those who earn high scores do differ on the criterion from those who earn lower scores.]

C 15.2 Validity of predictions from interest tests should be estimated separately at different levels of mental ability. DESIRABLE

C 16. Reports of validation studies should describe any conditions likely to affect the motivation of subjects for taking the test. ESSENTIAL

[Comment: If an ability test is to be used for employee selection, it should be validated using subjects who are candidates for employment and are therefore motivated to perform well. Under some testing conditions, a subject might try to "fake" his self-report of interests or personality; the controls used to discourage such faking should be reported.]

Concurrent Validity

All recommendations listed under predictive validity also apply to reports of concurrent validity, with the exception of C 5.

C 17. Reports of concurrent validity should be so described that the reader will not regard them as establishing predictive validity. ESSENTIAL

[Comment: The Minnesota Teacher Attitude Inventory is validated against contemporary teaching performance. This is reported under the general heading of "validity," and use of the test for selecting teachers or teacher-training candidates is recommended. The manual should point out that there have so far been no studies measuring entering students and observing them later on the job.]

C 17.1 For occupational tests where there are no longitudinal studies following subjects from the time of testing to the point where criterion information is available, validation data obtained by testing samples of employed persons should be presented. VERY DESIRABLE

[Comment: One such method of preliminary validation is to compare the distribution of scores for men in an occupation with those for men-in-general.]

C 17.11 If data from employed persons are used, evidence as to the effects of experience on interest inventory scores should be presented. ESSENTIAL

Construct Validity

Recommendations C 3–C 16 and D 5 apply to some reports of construct validity.

C 18. The manual should report all available information which will assist the user in determining what psychological attributes account for variance in test scores. ESSENTIAL

C 18.1 The manual should report correlations between the test and other tests which are better understood. VERY DESIRABLE

[Comment: It is desirable, for instance, to know the correlation of an "art aptitude" test for college freshmen with measures of general or verbal ability, and also with measures of skill in drawing. The interpretation of test scores would differ, depending on whether these correlations are high or low. On the other hand, it is clearly impractical to ask that the test author correlate his test with all prominent tests. It is especially valuable to know correlations of this test with other measures likely to be used in making decisions about the person tested.]

C 18.2 The manual should report the correlations of the test with other previously published and generally accepted measures of the same attributes. VERY DESIRABLE

[Comment: When a test is advanced as a measure of "general adjustment," its correlation with one or more other such measures should be reported. Similarly, if a test is advanced as a measure of "mechanical interest" or "introversion," its correlations with other measures of

these traits should be reported. The user can infer, from the size of such correlations, whether generalizations established on the older test can be expected to hold for the new one. Practical limitations will prevent the author from correlating his test with all competing tests. An example of good practice is the report, in the Thurstone Interest Schedule, of correlations with corresponding Kuder scores.]

C 18.3 If a test given with a time limit is to be interpreted as measuring a hypothetical psychological attribute, evidence should be presented concerning the effect of speed on test scores and on the correlation of scores with other variables. VERY DESIRABLE

C 18.4 If a test has been included in factorial studies which indicate the proportion of the test variance attributable to widely known reference factors, such information should be presented in the manual. DESIRABLE

C 19. The manual for a test which is used primarily to assess postulated attributes of the individual should outline the theory on which the test is based and organize whatever partial validity data there are to show in what way they support the theory. VERY DESIRABLE

D. Reliability

Reliability is a generic term referring to many types of evidence. The several types of reliability coefficient do not answer the same questions and should be carefully distinguished. We shall refer to a measure based on internal analysis of data obtained on a single trial of a test as a *coefficient of internal consistency.* The most prominent of these are the analysis of variance method (Kuder-Richardson, Hoyt) and the split-half method. A correlation between scores from two forms given at essentially the same time we shall refer to as a *coefficient of equivalence.* The correlation between test and retest, with an intervening period of time, is a *coefficient of stability.* Such a coefficient is also obtained when two forms of the test are given with an intervening period of time.

[Comment on projective tests: It is generally recognized that projective tests present even more than the usual difficulties in assessing reliability. It is not always clearly appropriate to demand internal consistency or stability and as yet equivalent forms for the most part do not exist. It seems reasonable, however, to require an assessment of stability for such instruments even though it is recognized in some instances that a low retest stability over a substantial period merely reflects true trait fluctuation and hence indicates good validity. Clinical practice rarely presumes that the inference from projective tests are to be applied on the very day the test is given. Realistically, we must recognize that pragmatic decisions are being made from test data which are meaningful only in terms of at least days, and usually weeks or months, of therapy and other procedures following the test administration. If a certain test result is empirically found to be highly unstable from day to day, this evidence casts doubt upon the utility of the test for most purposes even if that fluctuation might be explained by hypothesis of trait inconstancy.

This reasoning applies strictly only to the inferred dimensions, and not necessarily to the directly scored dimensions. If a personality variable is estimated from a complex of several test variables, and in such a way that rather different combinations of the test variables can

lead to the same value of the estimate, it is the temporal stability of the *estimate* which is subjected to the preceding requirement. But the burden of proof lies clearly upon the test manual. If component scores are unstable, it is then necessary to gather evidence regarding the degree to which estimates of the underlying personality dimension are stable during the interval for which they are intended to be used.]

D 1. The test manual should report such evidence of reliability as would permit the reader to judge whether scores are sufficiently dependable for the recommended uses of the test. If any of the necessary evidence has not been collected, the absence of such information should be noted. ESSENTIAL

D 1.1 Recommendation D 1 applies to every score, subscore, or combination of scores whose interpretation is suggested. ESSENTIAL

D 1.2 If differences between scores are to be interpreted or if the plotting of a profile is suggested, the manual should report the reliability of differences between scores. ESSENTIAL

D 1.21 If reliability of differences between an individual's scores is low, the manual should caution the user against interpreting profiles or score differences except as a source of preliminary information to be verified. ESSENTIAL

[Comment: The California Test of Mental Maturity reports reliability coefficients for the main scores and for scores on the major sections. Each section is further divided, the Spatial subtests, for example, including a group of items on Manipulation of Areas. By listing scores for such subsections on the profile sheet, the authors indirectly encourage interpretation of them. While supplementary material on the test mentions the low reliability of the subsections, the manual does not. It would be sounder practice to plot only those scores whose reliability is determined and reported in the manual.

The Watson-Glaser Critical Thinking Appraisal suggests that study of pupil performance on various types of items may enrich the interpretation. The manual adds this desirable caution:

"For a relatively small number of items such indices and special scores would not have high statistical reliability, and hence attention should be paid only to extreme deviates. For this reason norms for these special scores are not given and they are suggested only as an aid in helping students."

This paragraph illustrates how a manual may conform to the spirit of the Technical Recommendations even when some form of data is not provided in the manual. The statement would be improved if it were worded "such indices are probably unreliable" in the place of the present correct but euphemistic phrasing.]

D 1.3 One or more measures of reliability should be reported even when tests are recommended solely for empirical prediction of criteria. DESIRABLE

[Comment: The E. R. C. Stenographic Aptitude Test reports validity coefficients without also giving an estimate of reliability. For certain judgments such as the potential effect of lengthening the test information about reliability is required and should be available to the user.]

D 1.4 In connection with reliability measures, the manual should report whether the error of measurement varies at different score levels. If there is significant change in the error of measurement from level to level, this fact should be properly interpreted. VERY DESIRABLE

[Comment: Terman and Merrill point out that differences in IQ from Form L to Form M of the Revised Stanford-Binet Scale are much larger for IQ's above 100 than for low IQ's.

The California Test of Personality intentionally yields markedly skewed scores. This lowers the reliability coefficients from the value that might be attained with a normal distribution of raw scores, but reduces the error of identifying the most maladjusted cases. Here the most appropriate information on reliability would be the expected variation of percentile scores from trial to trial, reported separately for low and high scores.]

D 1.5 Reports of reliability studies should ordinarily be expressed in terms of: (*a*) the product-moment correlation coefficient; (*b*) another standard measure of relationship suitable to categorical judgments; or (*c*) the standard error of measurement. ESSENTIAL

[Comment: Chi square is not an adequate index of reliability for categorical judgments, since it reflects level of significance rather than magnitude of relationship.]

D 2. The manual should avoid any implication that reliability measures demonstrate the predictive or concurrent validity of the test. ESSENTIAL

[Comment: Properly interpreted reliability coefficients may support analysis of content or construct validity.]

D 3. In reports of reliability, procedures and sample should be described sufficiently for the reader to judge whether the evidence applies to the persons and problem with which he is concerned. ESSENTIAL

D 3.1 Evidence of reliability should be obtained under conditions like those in which the author recommends that the test be used. VERY DESIRABLE

[Comment: The maturity of the group, the variation in the group, and the attitude of the group toward the test should represent normal conditions of test use. For example, the reliability of a test to be used in selecting employees should be determined by testing applications for positions rather than by testing college students, or workers already employed.]

D 3.2 The reliability sample should be described in terms of any selective factors related to the variable being measured, usually including age, sex, and educational level. Number of cases of each type should be reported. ESSENTIAL

D 3.3 Appropriate measures of central tendency and variability of the test scores of the reliability sample should be reported. ESSENTIAL

D 3.31 If reliability coefficients are corrected for restriction of range, the nature of the correction should be made clear. The manual should also report the uncorrected coefficient, together with the standard deviation of the group tested and the standard deviation assumed for the corrected sample. In discussing such coefficients, emphasis should be placed on the one which refers to the degree of variation within which discrimination is normally required. ESSENTIAL

D 3.4 When a test is ordinarily required to make discriminations within a subclass of the total reliability sample, the reliability within each class should be investigated separately. If the coefficients differ, each separate coefficient should be reported. VERY DESIRABLE

[Comment: The Mechanical Reasoning section of the Differential Aptitude Tests has different reliability for boys and girls. The manual reports the reliability for each sex and grade.]

D 3.5 The manual should not imply that if some method had been used to determine reliability other than the one actually used, an appreciably higher coefficient would have been obtained. ESSENTIAL

Equivalence of Forms

D 4. If two forms of a test are made available, with both forms intended for possible use with the same subjects, the correlation between forms and information as to the equivalence of scores on the two forms should be reported. If the necessary evidence is not provided, the manual should warn the reader against assuming comparability. ESSENTIAL

D 4.1 Where two trials of a test are correlated to determine equivalence, the time between testings should be stated. ESSENTIAL (see also D 7)

D 4.2 Where the content of the test items can be described meaningfully, a comparative analysis of the forms is desirable to show how similar they are. DESIRABLE

Internal Consistency

D 5. If the manual suggests that a score is a measure of a generalized, homogeneous trait, evidence of internal consistency should be reported. ESSENTIAL

[Comment: Internal consistency is important if items are viewed as a sample from a relatively homogeneous universe, as in a test of addition with integers, or a test presumed to measure introversion. In a test which is regarded as a collection of diverse items, such as the Mooney Checklist, internal consistency is a minor consideration.]

D 5.1 When a test consists of separately scored parts or sections, the correlation between the parts or sections should be reported. ESSENTIAL

[Comment: Whether it is desirable or undesirable to have high subtest correlations depends on the nature and purpose of the test. Information on homogeneity or internal consistency may be relevant to the construct validity of the test.]

D 5.11 If the manual reports the correlation between a subtest and a total score, it should point out that part of this correlation is an artifact. ESSENTIAL

[Comment: Desirable practice is illustrated in the 1953 manual for the California Test of Personality.]

D 6. Coefficients of internal consistency should be determined by the split-half method or methods of the Kuder-Richardson type, if these can properly be used on the data under examination. Any other measure of internal consistency which the author wishes to report in addition should be carefully explained. ESSENTIAL

[Comment: There will no doubt be unusual circumstances where special coefficients give added information. There are grave dangers of giving unwarranted impressions, however, as is illustrated in the case of the Brainard Occupational Preference Inventory. This test yields a set of scores which are interpreted as a profile. The manual reports no information on the reliability of these scores, but does report a "total reliability" based on

a formula by Ghiselli. This reliability seems not to correspond to any score actually interpreted, and what it indicates about the value of this particular test is unclear without more discussion than the manual provides.

The original Kuder-Richardson formulas apply to a restricted case. Of those formulas, the one known as Number 20 is most satisfactory. A formula given by Hoyt, and others, has the same meaning but is more general in application.

Guttman has also suggested a "reproducibility" formula which relates to internal consistency. This index presents such special problems that it seems to have little suitability for test manuals.]

D 6.1 For time-limit tests, split-half or analysis of variance coefficients should never be reported unless: (*a*) the manual also reports evidence that speed of work has negligible influence on scores; or (*b*) the coefficient is based on the correlation between parts administered under separate time limits. ESSENTIAL

[Comment: Evidence of accuracy of measurement for highly speeded tests is properly obtained by retesting or testing with independent equivalent forms. If better evidence is not available, it is appropriate to use lower-bound formulas designed for estimating the internal consistency of speeded tests to determine the minimum coefficient.]

D 6.2 If several questions within a test are experimentally linked so that the reaction to one question influences the reaction to another, the entire group should be treated as an "item" in applying the split-half or analysis of variance methods. ESSENTIAL

[Comment: In a reading test, several questions about the same paragraph are ordinarily experimentally dependent. All of these questions should be placed in the same half-test in using the split-half method. In the Kuder-Richardson method, the score on the group of questions should be treated as an "item" score.]

D 6.3 If a test can be divided into sets of items of different content, internal consistency should be determined by procedures designed for such tests. VERY DESIRABLE

[Comment: One such procedure is the division of the test into "parallel" rather than random half-tests. Another procedure is to apply the Jackson-Ferguson "battery reliability" formula.]

Stability

D 7. The manual should indicate what degree of stability of scores may be expected if a test is repeated after time has elapsed. If such evidence is not presented, the absence of information regarding stability should be noted. ESSENTIAL

[Comment: Most educational and psychological tests measure qualities which are presumed to be stable for some time, unless training or specified experiences intervene. Stability is not always desirable. A measure of interests in childhood and adolescence which is highly stable would not be sensitive to developmental changes.]

D 7.1 Stability of scores should be determined by administering the test to the same group at different times. The manual should report changes in mean score as well as the correlation between the two sets of scores. ESSENTIAL

D 7.11 If a test result is reported in terms of pass-fail or some other categorical classification, stability should be reported in terms of proportion of altered classifications on retest. VERY DESIRABLE

D 7.12 In determining stability

of scores by repeated testing, other precautions such as giving alternate forms of the test should be used to minimize recall of specific answers, especially if the time-interval is not long enough to assure forgetting. VERY DESIRABLE

D 7.13 In reporting a coefficient of stability, the manual should describe the experience or education of the group between testings, if this would be expected to affect test scores. ESSENTIAL

D 7.2 For tests of interest and ability intended for use prior to adulthood, the coefficient of stability should correlate scores obtained at one particular age with scores at some later significant age. Coefficients should be reported separately for different ages at first test and for different periods of intervening time. ESSENTIAL

E. Administration and Scoring

E 1. The directions for administration should be presented with sufficient clarity that the test user can duplicate the administrative conditions under which the norms and data on reliability and validity were obtained. ESSENTIAL

E 1.1 The published directions should be complete enough so that people tested will understand the task in the way the author intended. ESSENTIAL

[Comment: If, for example, in a personality inventory, it is intended that subjects give the first response that occurs to them, this should be made clear in the directions for administration. Directions for interest inventories should specify whether the person is to mark what he would ideally like to do, or whether he is also to consider the probability that he would have the opportunity and ability to do them. Likewise, the directions should specify whether the person is to mark those things he would wish to do or does occasionally, or only those things he would like to do or does regularly.]

E 1.2 If expansion or elaboration of instructions, giving of hints, etc., is permitted, the conditions for it should be clearly stated either in the form of general rules or by giving numerous examples, or both. VERY DESIRABLE

E 1.21 If the examiner is allowed freedom and judgment in elaborating instructions or giving samples, empirical data should be presented regarding the effect of variation in examiner procedures upon scores. If empirical data on the effect of variation in examiner procedure are not available, this fact should be explicitly stated and the user warned that the effects of such variation are unknown. ESSENTIAL

E 1.3 If the test under consideration is of a type where previous experience demonstrates that subjects are likely to present an unrealistic picture of themselves, the manual should give evidence regarding the extent to which such distortion may affect scores. ESSENTIAL

[Comment: Such evidence is ordinarily to be provided by measuring the shift of scores when the test is administered in different situations (e.g., pre-employment and postemployment) or with instructions intended to induce different sets. This problem is especially acute for personality and interest inventories and projective techniques.]

E 1.31 If the test is provided with a verification key or key to correct for inappropriate test-taking attitudes.

evidence that this key performs its function should be provided. ESSENTIAL

E 2. Where subjective processes enter into the scoring of the test, evidence on degree of agreement between independent scorings should be presented. If such evidence is not provided, the manual should draw attention to scorer error as a possible source of error of measurement.

[Comment: With projective tests, the role of interscorer agreement in the actual *classification* of raw response data is more crucial than in the case of a test where an "error in scoring" means a clerical error or something close to that. Interscorer agreement is not a demonstration of reliability in the usual sense, or a substitute for it. Interscorer agreement deals solely with the objectivity of classifying the behavior sampled from subjects, and is, therefore, directed at a condition on the part of the judge's behavior that is *necessary* for "reliability." Interscorer consistency is obviously not a *sufficient* condition, since it cannot possibly give information regarding the adequacy of that behavior *as* a sample from the subject.]

E 2.1 The bases for scoring and the procedure for training the scorers should be presented in sufficient detail to permit other scorers to reach the degree of agreement reported in studies of scorer agreement given in the manual. VERY DESIRABLE

[Comment: One desirable practice is to present a list of the commoner responses or response categories with their scoring indicated.]

E 2.11 If persons having various degrees of supervised training are expected to score the test, studies of the interscorer agreement at each skill level should be presented. DESIRABLE

E 2.2 If reliability of scoring is low, the manual should caution the user against interpreting combinations of such scores. ESSENTIAL

[Comment: Combinations such as ratios generally will be even less reliable than the component scores.]

F. Scales and Norms

F 1. Scales used for reporting scores should be such as to increase the likelihood of accurate interpretation and emphasis by test interpreter and subject. ESSENTIAL

[Comment: Scales in which test scores are reported are extremely varied. Raw scores are used. Relative scores are used. Scales purporting to represent equal intervals with respect to some external dimension (such as age) are used. And so on. It is unwise to discourage the development of new scaling methods by insisting on one form of reporting. On the other hand, many different systems are now used which have no logical advantage, one over the other. Recommendations below that the number of systems now used be reduced to a few with which testers can become familiar, are not intended to discourage the use of unique scales for special problems. Suggestions as to preferable scales for general reporting are not intended to restrict use of other scales in research studies.]

F 2. Where there is no compelling advantage to be obtained by reporting scores in some other form, the manual should suggest reporting scores in terms of percentile equivalents or standard scores. VERY DESIRABLE

[Comment: Professional opinion is divided on the question whether mental test scores should be reported in terms of some theoretical growth scale, such as the intelligence quotient or the Heinis index. Thus, a test developer who has ration-

ale for such scales as these should use them if he regards them as especially adequate.

On the other hand, there is no theoretical justification for scoring mental tests in terms of an "IQ" which is not derived in terms of the theory underlying the Binet IQ and which has different statistical properties than the IQ does. Standard or percentile scores would be preferable to arbitrarily defined IQ scales such as are used in the Otis Gamma and Wechsler-Bellevue tests.

Strong recommends that Vocational Interest Blank scores be converted into letter grades where "A" indicates that at least two-thirds of the criterion group equaled or exceeded a given score, etc. He bases this recommendation on the ground that finer score discriminations would lead only to unwarranted attempts at finer interpretative discrimination.]

F 2.1 If grade norms are provided, tables for converting scores to percentiles (or standard scores) within each grade should also be provided. ESSENTIAL

[Comment: At the high school level, norms within courses (e.g., second year Spanish) may be more appropriate than norms within grades.]

F 3. Standard scores obtained by transforming scores so that they have a normal distribution and a fixed mean and standard deviation should in general be used in preference to other derived scores. For some tests, there may be a substantial reason to choose some other type of derived score. VERY DESIRABLE

F 3.1 If a two-digit standard score system is used, the mean of that system should be 50 and the standard deviation 10. DESIRABLE

F 3.2 If a one-digit standard score system is used, the mean of the system should be 5 and the standard deviation 2 (as in stanines). DESIRABLE

[Comment: The foregoing are proposed as ways of standardizing practice among test developers. It is expected that institutions with established systems, such as the College Board Scale, with mean at 500, will often retain them as suited to their purposes.]

F 3.3 Where percentile scores are to be plotted on a profile sheet, the profile sheet should be based on the normal probability scale. VERY DESIRABLE

F 4. Local norms are more important for many uses of tests than published norms. In such cases the manual should suggest appropriate emphasis on local norms. VERY DESIRABLE

[Comment: The Cooperative Dictionary Test manual precedes its presentation of norms with a discussion urging schools to prepare local norms and explaining their advantages over the published norms with respect to this test. Many achievement tests, clinical tests, and tests used for vocational guidance might well present a similar statement.]

F 5. Except where the primary use of a test is to compare individuals with their own local group, norms should be published at the time of release of the test for operational use. ESSENTIAL

[Comment: The Thurstone Interest Schedule provides a profile of 20 raw scores. Because each field is based on the same number of items, norms are said to be unnecessary. Yet a change of items in any group would make that category more or less preferred. Hence, to know whether a high score reflects this individual's interests, or only that these items are popular with everyone, the user must consult a set of norms. Judg-

ment in terms of raw scores could be made only if by some unusual method it could be demonstrated that the items in each category are a representative sample of that field.]

F 5.1 Even though a test is used primarily with local norms, the manual should give some norms to aid the interpreter who lacks local norms. DESIRABLE

F 6. Norms should report the distribution of scores in an appropriate reference group or groups. ESSENTIAL

F 6.1 Unless they can be readily inferred from the table of norms, measures of central tendency and variability of each distribution should be given. ESSENTIAL

F 6.2 If the distribution in the norm group is not essentially normal, some form of percentile table should be provided. ESSENTIAL

F 6.3 In addition to norms, tables showing what expectation a person with a given test score has of attaining or exceeding some relevant criterion score should be given where possible. Conversion tables translating test scores into proficiency levels should be given when proficiency can be described on a meaningful absolute scale. DESIRABLE

F 7. Norms should refer to defined and clearly described populations. These populations should be the groups to whom users of the test will ordinarily wish to compare the persons tested. ESSENTIAL

[Comment: Intelligence tests designed for use with elementary school children might well present norms by grade-groups as well as by chronological age-groups.

For occupational inventories, norms based on men who have entered specific occupations should be developed, except where cutting scores or regression formulas are provided for predicting occupational criteria.

The manual should point out that a person who has a high degree of interest in a curriculum or occupation, when compared to men-in-general, will generally have a much lower degree of interest when compared with persons actually engaged in that field.

Thus a *high* percentile score on the Kuder mechanical scale, in which the examinee is compared with men-in-general, may be equivalent to a *low* percentile when the examinee is compared with auto mechanics.]

F 7.1 The manual should report the method of sampling within the population, and should discuss any probable bias within the sample. ESSENTIAL

F 7.11 Norms should be based on a well-planned sample rather than on data collected primarily on the basis of availability. VERY DESIRABLE

[Comment: Occupational and educational test norms have often been based on scattered groups of test papers, and authors sometimes request that all users mail in results for use in subsequent reports of norms. Distributions so obtained will contain unknown biases. Hence, the methods for obtaining the samples should be clearly described, as in Strong's manual, and whenever possible, samples should be stratified to remove some of the bias. Planned samples will give more dependable norms, however, since stratification cannot remove all sampling error.]

F 7.2 The number of cases on which the norms are based should be reported. ESSENTIAL

F 7.21 If the sample on which

norms are based is small or otherwise undependable, the user should be explicitly cautioned regarding this. ESSENTIAL

[Comment: In addition to general high school and college norms based on substantial samples, medians and ranges in small, special groups are reported for the Watson-Glaser Critical Thinking Appraisal. Since these samples vary from 10 cases to 65 cases, the ranges and medians are highly unstable. This manual should report quartiles in preference to ranges because quartiles are more stable. This manual should warn the reader as to the fallibility of estimates from these special samples.]

F 7.3 The manual should report whether scores differ for groups differing on age, sex, amount of training, and other equally important variables. ESSENTIAL

F 7.31 If appreciable differences between groups exist, and if a person would ordinarily be compared with a subgroup rather than with a random sample of persons, then separate norm tables should be provided in the manual for each group. ESSENTIAL

[Comment: An example of unusually excellent practice is the norms for the Minnesota Teacher Attitude Inventory. Here norms are based on teachers separated by levels of experience, amount of training, and type of position. The teachers were obtained by a planned sample. The manual discusses differences between sex groups but does not present separate norms, as the decision to employ a particular man teacher rather than a woman would be based on the raw score of each, rather than upon their standings within their sex group.

Norms for interest inventories should be prepared separately for student examinees having different levels of general academic ability unless there is evidence that scores have no relation to ability.]

F 7.32 When the total amount of scorable behavior is allowed by the task to vary, separate norms on the various scored variables should be presented for different levels of total response. VERY DESIRABLE

F 7.33 If the standardizing sample is too small to permit calculation of separate norms on scored variables at different levels of total response, the correlation of each of these with response level must be presented. ESSENTIAL

F 7.34 If correlation data suggest that the dependence of scores on total responsiveness is nonlinear, this should be explicitly stated and the user warned that linear corrections, prorating, or computing of percentages are inappropriate procedures. ESSENTIAL

F 7.35 If data are insufficient to determine the nature of the dependence of the several scores upon responsivity (such as linearity, array scatter), this lack of information should be explicitly mentioned and the possible dangers in interpretation should be stressed. ESSENTIAL

F 7.4 If conditions affecting test scores are expected to change as time elapses, periodic review of norms is required. VERY DESIRABLE

F 7.5 Some profile sheets record, side by side, scores from tests so standardized that different scores compare the person to different norm groups. Profiles of this type should be recommended for use only where tests are intended to assess or predict the person's standing in different situations, where he competes with

the different groups. Where such mixed scales are compared, the fact that the norm groups differ should be made clear on the profile sheet. VERY DESIRABLE

F 7.6 The description of the norm group should be sufficiently complete so that the user can judge whether his case falls within the population represented by the norm group. The description should include number of cases, classified by relevant variables such as age, sex, educational status, etc. ESSENTIAL.

F 7.7 The conditions under which normative data were obtained should be reported. The conditions of testing, including the purpose of the subjects in taking the test, should be reported. ESSENTIAL

[Comment: Some tests are standardized on job applicant groups, others on groups which have requested vocational guidance, and still others on groups which realized they were "guinea pigs." Motivation for taking tests, test-taking attitudes, abilities, and personality characteristics possibly differ on all of these groups.]